Paper 3140-2015

An Application of the Cox Proportional Hazards Model to the Construction of Objective Vintages for Credit in Financial Institutions, Using PROC PHREG

Iván Darío Atehortua Rojas, Banco Colpatria – Scotia Bank, Bogotá Colombia

ABSTRACT

Retail segment in a Bank is a major source of incomes. Especially for a bank with a share in the market of 17% in number of credit cards, the biggest in Colombia. This quantity of lending applications force the bank to use analytic and statistical tools to do an initial selection of good customers. And then analyze each of those applications that had the lowest probability of default and will imply secure incomes to the organization.

Logit models are used for the construction of acquisition models. Where the objective is to estimate the probability that in a certain fixed period of time, usually 12 months, the customer falls into default. Besides, to know the probability that a customer falls into default in different time periods (e.g. one, two, … , n months after the entailment) is necessary for credit risk vintage analysis.

Hence Cox Proportional Hazards Model becomes important. Because with this model, probabilities are more predictive. Then the construction of a target risk curve to know in the short term if new credits have a good behavior will be more accurate.

The construction of target vintages using the Cox model will generate alerts in increasing of days past due in a shorter time, so the mitigation measures can be applied one or two months earlier than currently and this can reduce the losses by 100 bps in the new vintages. This paper makes the estimation of a proportional hazard model of Cox and compare the results with a logit model for a specific product of the Bank. Additionally, we will estimate the objective vintage for the product.

INTRODUCTION

One of the most studied topics on credit risk industry is knowing the probability of default of a customer (PD). There are several methodologies to estimate the PD in the future. One of those is to model the probability of been a bad customer, taking into consideration current and antique variables for each customer. To use various time periods is will also improve the quality of the estimation.

In general to define a customer as good or bad, a common methodology is to calculate the maximum dpd for a customer in the performance window (period of time where costumers are rated as or good or bad). Furthermore is necessary to build up a bad customer definition that is a break point from which the client will be considered as a bad client to the bank. (common definitions: 30dpd, 60dpd, 90dpd…).

Nevertheless the described methodology do not have into account the differences between a customer with a maximum dpd of 60 in the month 3 since origination than another with the same dpd but in the month 12 since origination. This additional information is captured by cox models and differentiate them from most of the traditional models used in credit risk as the logit model.

These paper is dived into 4 sections. First section expose concepts used in the model. Second and third sections explain the data and methodology used for modelling. Finally current results and comparison with a logit model model is exposed.

GENERAL CONCEPTS

Vintages

Vintages in a credit risk context shows the behavior of a portfolio. In scoring models specifically, a vintage indicates which percent of clients or balances of clients fall into default in different periods of time after the credit is originated.

Rejected Inference

When a model is in production, that model causes some rejects because the client has a low score. If we want to replace that model, we have to use the behavior of clients after they were approved to construct the objective variable (in the performance window). However, if we were using another model and we want to change it, the caused rejects of that model will not have behavior and we will have to infer their behavior and to include the rejects in the model because if we do not do it then we will cause a bias.

PSI

Population Stability Index (PSI), measures the shift in applicant score distribution. It is how our current populations have changed respect the construction population. Higher PSI values indicate a larger shift in the distribution. (Mays, 2001)

KS

Kolmogorov-Smirnov statistic (KS) is the maximum difference between the cumulative percent good distribution and the cumulative percent bad distribution. Loans are ranked from low to high by score then divided up into score ranges. When KS is high the model discriminate better. (Mays, 2001)

DATABASE DESCRIPTION

Used data base contains 42.322 for one year through the door applications for one of the credit cards products offered by the Bank. 2.829 of those are rejected applications with previous models, for which we will do a rejected inference analysis. Additionally 4.849 accepted applications were analyzed during last two months to calculate the population stability index.

46 variables were analyzed which include financial sector behavior variables and socio-demographic variables.

METHODOLOGY

Survival Analysis

In survival analysis, dependent variable is failure time, these is the time until studied event occurs. Main characteristic of survival models is the presence of the censoring which is generated once the time during the subject is analyzed is ended and the predicted event didn't occurs. (Colosimo, 2007).

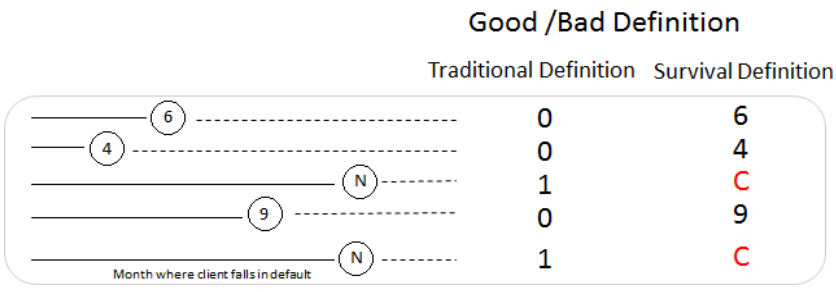

Figure 1. Good Bad Definition in survival analysis

In the **Error! Reference source not found.**, we can see how to construct the objective variable. While in logit models we made a binomial variable where 1 represents the event where client did not fall into default and 0 if he fell, in a survival model the objective variable contains the month where the client falls into default. For example the number 6 in the first row shows that the client falls into default in the sixth month. On the other hand, when the client does not fall into default, the event is censored as in the row three in the figure.

Survivor Function

One of the principal functions used in survival analysis is the Survivor Function. Let T be a not negative random variable than represents the time to the failure, The survivor function, is defined as the probability of a observation of does not fall into default before a time t. It is:

$$(1)\ S_t = P(T \geq t)$$

In consequence, the function $F_t = 1 - S_t$ represents the probability of a client of falls into default before the time t. Functions S_t and F_t are very important in survivor analysis and in Risk Analysis, F_t shows the probability of each client of falls into default in each month after the origination. This function is very important because if F_t shows a rapid growth, the risk and provisions to growth fast. F_t is the principal advantage of the survivor models over logit and other models.

Another important function in survivor analysis is the failure rate which is defined as the probability that the failure occur in a time interval given that it does not occur before the time interval divided by the interval length:

$$(2)\ \lambda_t = \frac{S_t - S_{t+\Delta t}}{\Delta t * S_t}$$

The next relations are demonstrable:

$$(3)\ S_t = e^{-\Lambda(t)} = e^{-\int_0^t \lambda(\mu) d\mu}$$

This relation is important because with Proc PHREG we obtain the $-\Lambda(t)$ function known as the cumulated failure rate function or hazard function.

The Cox Proportional Model

When we want to estimate the functions S_t, F_t or $\Lambda(t)$ for each client using his behavior and sociodemographic variables, the Cox Proportional Model is one of the most used models.

Suppose we want to compare the expected time of fail of two groups. Where rate failure function for each group is λ_1 and λ_2, Cox proportional risk model is applicable if:

$$(4)\ \frac{\lambda_1(t)}{\lambda_2(t)} = K$$

Where, relative risk between two groups is constant for every time period. Thus when we have these relationship, it can be specified as:

$$(5)\ \lambda(t; Z_i) = \lambda_0(t) e^{Z_i' B}$$

Where $\lambda_0(t)$ is the failure rate function of a base group, Z_i are the values of the dependent variables of the i customer, and B the estimated coefficients vector of the model. Furthermore,

$$(6)\ S(t; Z_i) = S_0(t)^{\exp(Z_i' B)}$$

Where $S(t; Z_i)$ is the survival function of the i^{th} customer in the period of time t and $S_0(t)$ corresponds to the survival function of a base customer in the period of time t.

KGB Model

As the replaced model is for acquisition. Rejected inference must be done to identify customers rejected by the previous model so they can be included in the modeling sample. This model is named as Known Good-Bad Model (KGB). Used methodology for constructing a KGB model is similar to the one used for the

estimation of the customer model which include reject inference customers also named as All Good-Bad Model (AGB).

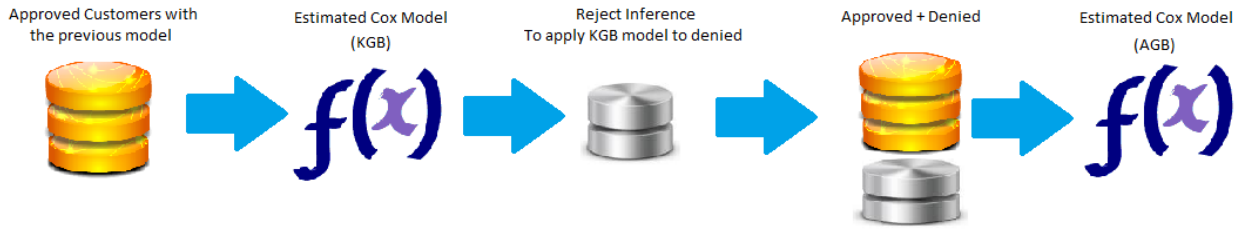

Figure 2. Reject Inference process

For a model were the dependent variable follows a binomial distribution. Once estimated the KGB model over the approved population a decile based score card can be done. This is dividing in deciles the obtained probability and calculating for each decile the percentage of bad customers.

Afterwards, rejected base must be rated and grouped in the same deciles ranks obtained before. Finally, for each customer a binomial distributed random variable is generated with probability of P_{decile} equal to the percentage of bads of each decile calculated over approved.

For a Cox Proportional Model, process is more complex because dependent variable do not follow a binomial distribution. To solve this issue a random variable was simulated.

	Probability of Fall in each month												
Z(i)B Rank	1	2	3	4	5	6	7	8	9	10	11	12	Censored
1	0.01%	0.06%	0.11%	0.43%	0.39%	0.33%	0.45%	0.42%	0.33%	0.62%	0.41%	0.39%	96.03%
2	0.01%	0.03%	0.19%	0.33%	0.49%	0.56%	0.64%	0.71%	0.74%	0.72%	0.67%	0.86%	94.06%
3	0.00%	0.01%	0.23%	0.40%	0.71%	0.53%	0.67%	0.92%	0.86%	0.89%	0.78%	0.83%	93.16%
4	0.00%	0.00%	0.21%	0.40%	0.61%	0.69%	0.93%	1.07%	1.01%	1.20%	1.00%	0.99%	91.88%
5	0.00%	0.01%	0.30%	0.58%	0.76%	0.84%	1.33%	1.65%	1.69%	1.61%	1.28%	1.51%	88.44%

Table 1. Probabilities used to reject inference.

In Table 1. there are the probabilities calculated for each month. That probabilities are monotonic in months and risk ranks. Finally, we generate a random variable with this distribution and depending the risk rank of each client, we assign a month of failure. Once we have ranked the reject base, we construct the AGB database using approvals + rejects.

RESULTS

Table 2 contains the results of estimating the cox proportional hazards model for the AGB population.

Analysis of Maximum Likelihood Estimates								
Parameter		Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Seniority (Months) Ref: >= 24	< 4	-0.18289	0.04324	17.892	<.0001	0.833	0.765	0.907
	4 - 11	-0.33698	0.05585	36.4002	<.0001	0.714	0.64	0.797
	12 - 23	-0.37643	0.07472	25.3784	<.0001	0.686	0.593	0.795
Client with mortgage	No	0.61366	0.09302	43.5217	<.0001	1.847	1.539	2.217
Number of Bureau inquiries in the last 6 months Ref: >=5	< 4	-0.80338	0.04966	261.7702	<.0001	0.448	0.406	0.494
	4 - 5	-0.43151	0.05407	63.6783	<.0001	0.65	0.584	0.722
Marital status Ref: Married	Common law marriage	0.21628	0.05792	13.9451	0.0002	1.241	1.108	1.391
	Single	0.24711	0.05961	17.1834	<.0001	1.28	1.139	1.439
	divorced / widowed	0.37029	0.07895	21.9963	<.0001	1.448	1.241	1.691
Age Ref: < 23	23 - 30	-0.12067	0.0553	4.7609	0.0291	0.886	0.795	0.988
	31 - 54	-0.31795	0.05824	29.7999	<.0001	0.728	0.649	0.816
	>= 55	-0.4872	0.0961	25.7007	<.0001	0.614	0.509	0.742
Max seniority Credit Cards Ref: >= 24	without cards	0.34471	0.06035	32.6212	<.0001	1.412	1.254	1.589
	1 - 12	0.43166	0.07227	35.6781	<.0001	1.54	1.336	1.774
	13 - 24	0.24192	0.08351	8.3916	0.0038	1.274	1.081	1.5
Kind of residence Ref: Own	Family	0.34255	0.05293	41.8861	<.0001	1.409	1.27	1.563
	Leased	0.41335	0.05889	49.2657	<.0001	1.512	1.347	1.697

Table 2. Estimated parameters for de AGB Model

Table 2 shows that risk increases in risk profiles, for example the model penalize clients with age < 23.

With PHREG, we can estimate the survivor function for the base line segment (Clients with reference categories in Table 2). In this case, results are:

Month	S(t)
0	1.000000
1	0.999967
2	0.999821
3	0.998325
4	0.995333
5	0.991098
6	0.986799

Month	S(t)
7	0.981092
8	0.974401
9	0.967782
10	0.960519
11	0.954680
12	0.947947

Table 3. Estimated Survival Function for the reference category.

Using the estimated survival function for the reference category ($S_0(t)$), we can find the survival functions for all our clients using the formula (6). Next code shows how to construct the survival functions and corresponding confidence intervals:

```
%let S0AGB_0=    1.0;
%let S0AGB_1=    0.999967458;
%let S0AGB_2=    0.999821026;
%let S0AGB_3=    0.998325132;
%let S0AGB_4=    0.995332549;
%let S0AGB_5=    0.991098074;
%let S0AGB_6=    0.986798558;
%let S0AGB_7=    0.981091797;
%let S0AGB_8=    0.974400651;
```

```
%let S0AGB_9=    0.967782408;
%let S0AGB_10=   0.960518766;
%let S0AGB_11=   0.954680249;
%let S0AGB_12=   0.947947107;


%MACRO Si_AGB(BASE);
data &BASE;
set &BASE;
%DO I = 0 %TO 12;
      S_&I.=&&S0AGB_&I.**EXP(XB_AGB);
      LI_S_&I.=&&S0AGB_&I.**EXP(IC_LI_XB);
      LS_S_&I.=&&S0AGB_&I.**EXP(IC_LS_XB);
      VINTAGE_&I.=1-S_&I.;
      LI_VINTAGE_&I.=1-LI_S_&I.;
      LS_VINTAGE_&I.=1-LS_S_&I.;
%END;
RUN;
%MEND();

%Si_AGB(Base_AGB);
```

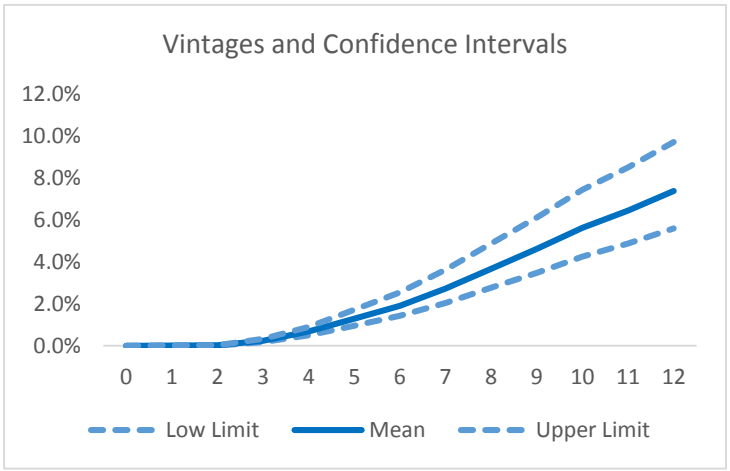

Figure 3. Vintage of clients

Figure 3 shows the mean of the estimated vintages ($F_t = 1 - S_t$) for the AGB population. This is the principal advantage of the cox proportional hazards model over a logit model. In the curve you can see that probability of fall into default start to growth from the third month. Also the curve shows that risk is ever increasing from month 3, it shows that we should take a longer performance window. Using the same methodology we can construct and compare vintages for different clusters:

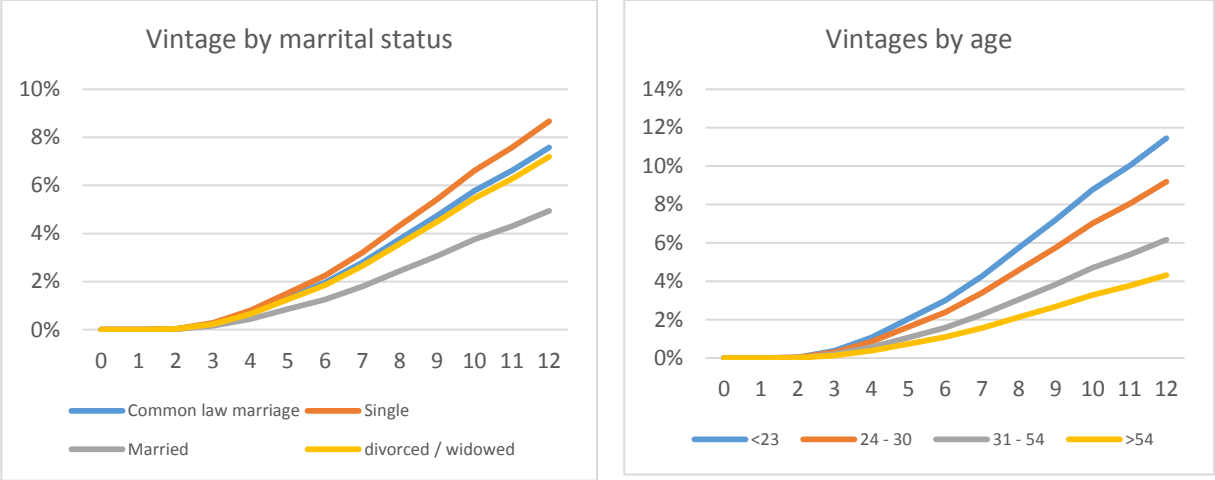

Figure 4. Vintages by different inputs

Finally, we compare the result of the estimated Cox proportional hazards model with a logistic model estimated with the same variables and results are very similar. For compare the two models, we use the response variable of the logit model and discriminatory power is very similar:

	KS	ROC
Logit Model	24.4%	66.3%
Cox Proportional Hazards Model	24.4%	66.2%

Stability is the same in the two models because we have the same variables. In this case, the PSI is 8.6%.

CONCLUSION

Proportional Hazards Models are a good alternative in origination models to logit models. The discriminatory power is very similar and we have an advantage to construct the vintages which are very used in credit risk.

REFERENCES

Anderson, Raymond. 2007. *The Credit Scoring Toolkit*. Oxford University Press.

Colosimo, Enrico Antonio. 2007. *Análise de sobrevivencia aplicada*. Departamento de Estadística UFMG.

Mays Elizabeth. et al. 2002. *Hand book of credit scoring*. American Management Association.

Montrichard, Derek. 2008. *Reject Inference Methodologies in Credit Risk Modeling*. Available at http://analytics.ncsu.edu/sesug/2008/ST-160.pdf

Thomas Lyn C. et al. 2002. *Credit Scoring and its applications*. Society for industrial and applied mathematics.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Iván Darío Atehortua Rojas
Banco Colpatria – Scotia Bank, Bogotá Colombia
+57 1 7456300 Ext. 3165
ivandarioate@gmail.com